Dynamic Days Asia Pacific 13, Kyoto Japan, 5 July 2024

Asymptotic Edge of Chaos as Guiding Principle for Neural Network Training

Ling Feng

Complex Systems Group, Institute of High Performance Computing, A*STAR Singapore

Department of Physics, National University of Singapore

Collaboration with

Lin Zhang (former PhD student)

Kan Chen

Choy Heng Lai

Zhang, Lin, et al., *International Journal of Artificial Intelligence and Robotics Research* 1.01 (2024).

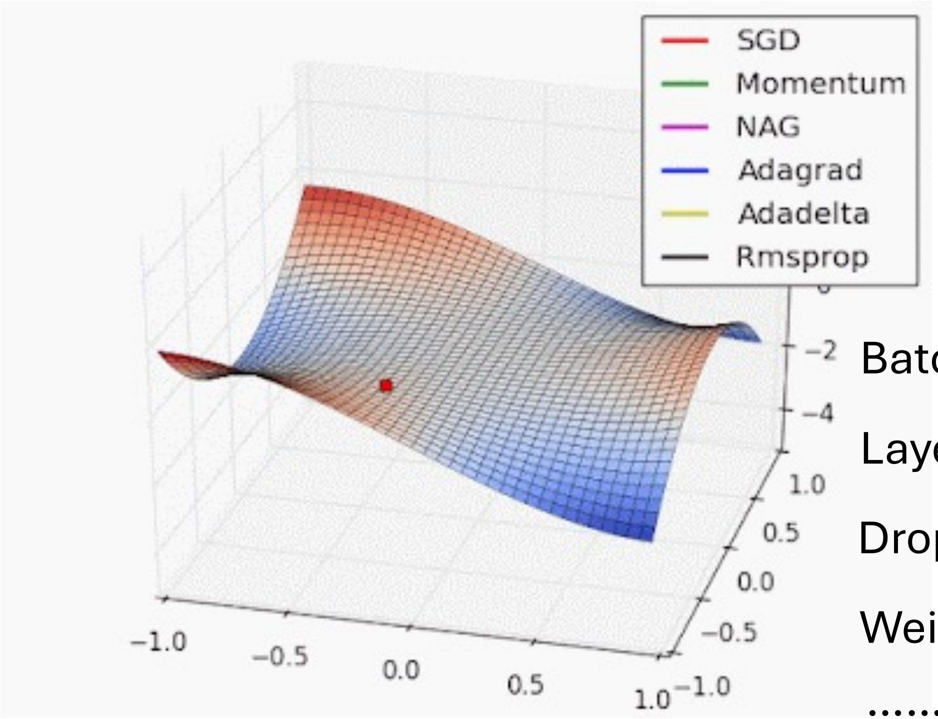
Background 1: Tuning AI training algorithms
- example: stochastic gradient descent (SGD) w. momentum

W_0 : Weights initialization (where to start training) — Starting temperature

η : Learning rate (Step size in gradient) — Stirring speed

B : Batch size (# of samples in each gradient computation) — Amount of ingredient to put in each time

1-α : Momentum (memory of previous gradients)



Batch Norm	Salt
Layer Norm	Pepper
Drop out	Spice
Weight decay	
……	……

https://www.denizyuret.com/2015/03/alec-radfords-animations-for.html

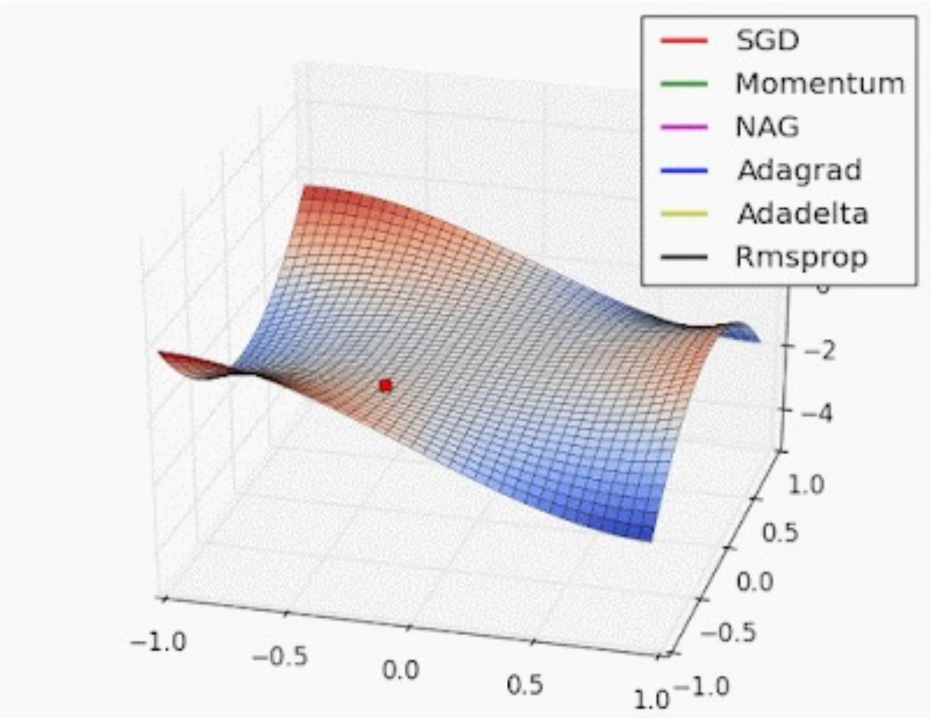
Details of SGD with momentum

W_0 : Weights initialization (where to start training)

η : Learning rate (Step size in gradient)

B : Batch size (# of samples in each gradient computation)

1-α : Momentum (memory of previous gradients)



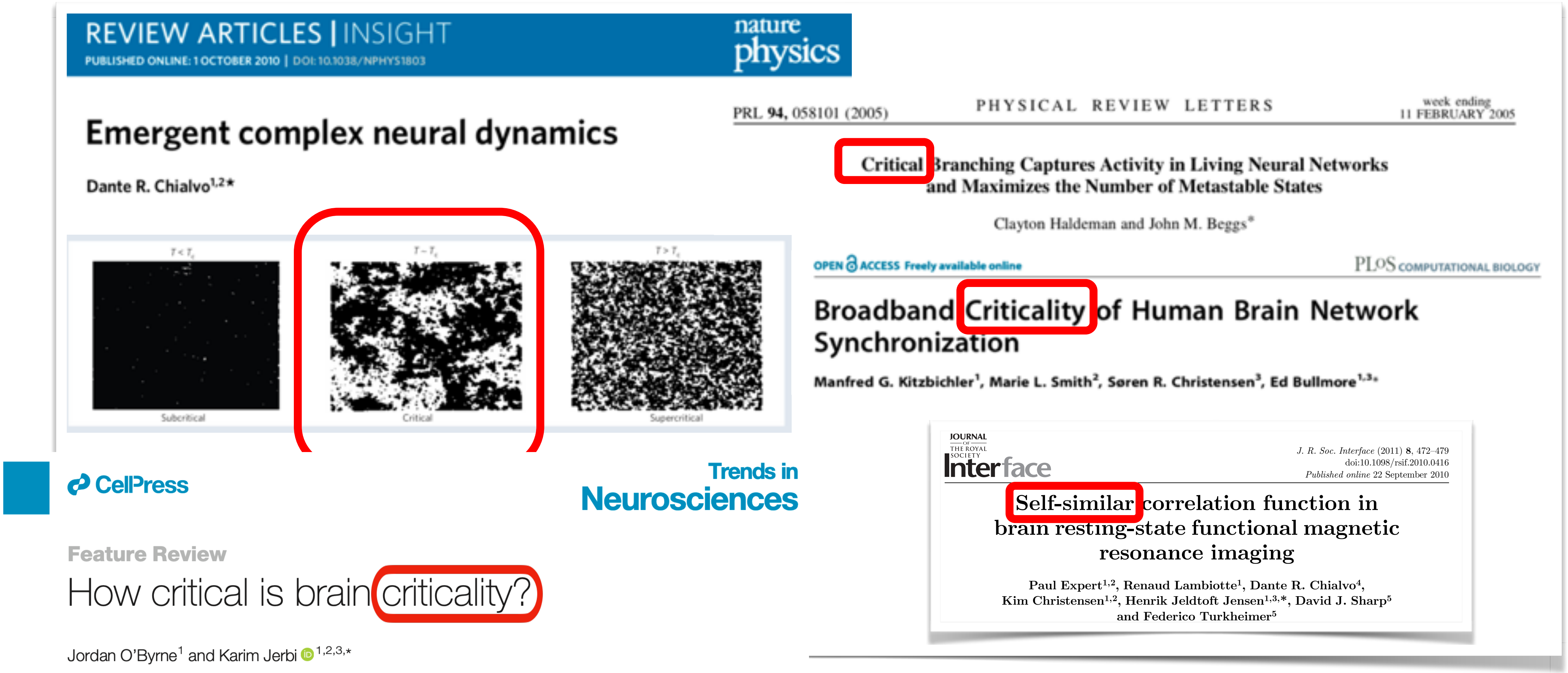
Adjustment to NN weights

$$\Delta w_t = v_t = \alpha v_{t-1} - \eta g_t$$

Gradient for backpropagation (Stochasticity depends on B)

Generally trial-and-error to determine these hyperparameters

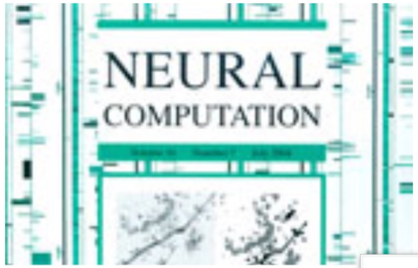
Background 2: Edge of chaos in NN
- healthy biological neural networks



Background 2: Edge of chaos in NN
- artificial neural networks computation

Information processing in echo state networks at the edge of chaos

Joschka Boedecker · Oliver Obst · Joseph T. Lizier · N. Michael Mayer · Minoru Asada



Real-Time Computation at the Edge of Chaos in Recurrent Neural Networks

Nils Bertschinger and Thomas Natschläger

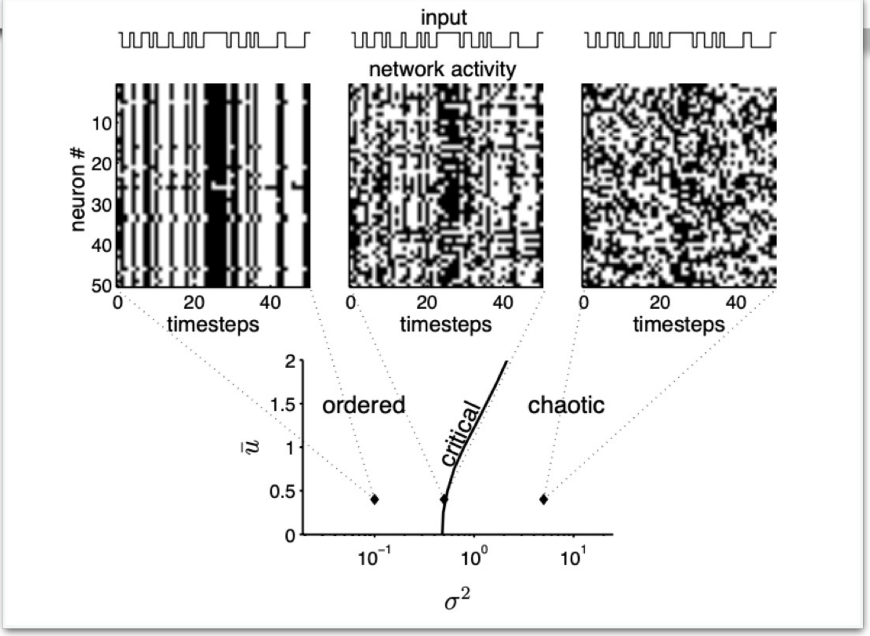
ELSEVIER

Neural Networks
Volume 20, Issue 3, April 2007, Pages 323-334

2007 Special Issue

Edge of chaos and prediction of computational performance for neural circuit models

Robert Legenstein, Wolfgang Maass



Background 2: Edge of chaos in NN
- deep neural networks (feedforward) performance

Residual Network

Convolutional Net (CNN)

DenseNet

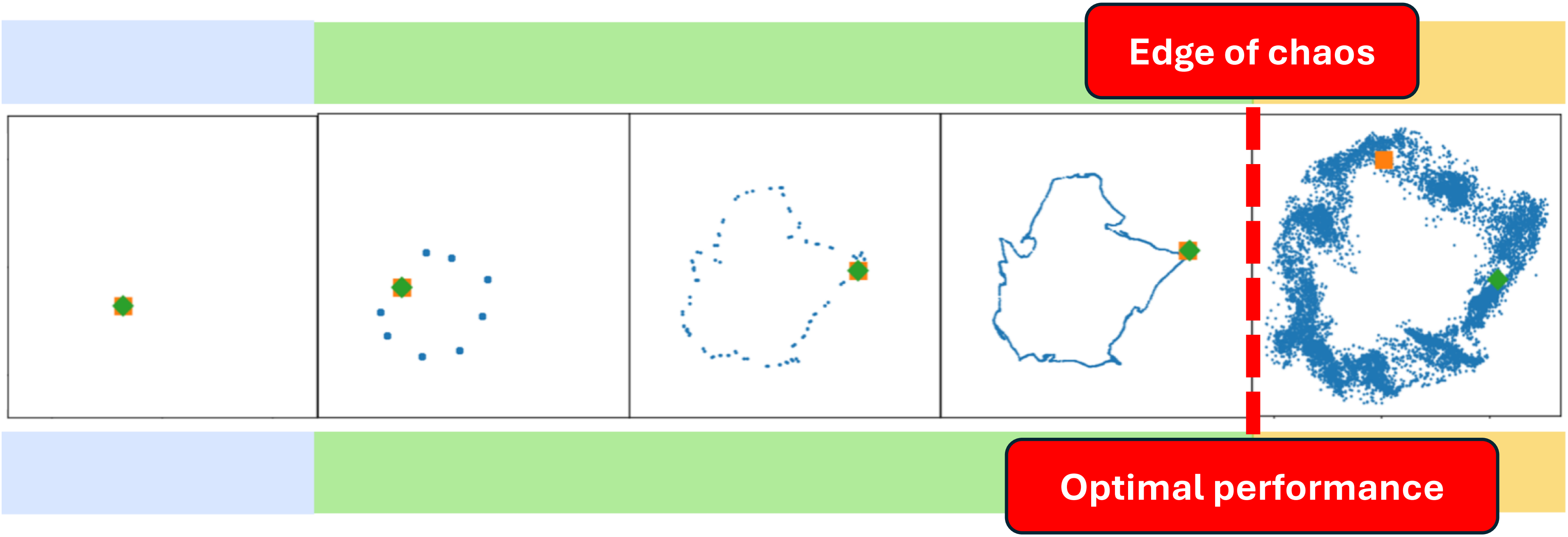
Batch Norm

Dropout

Weight Decay

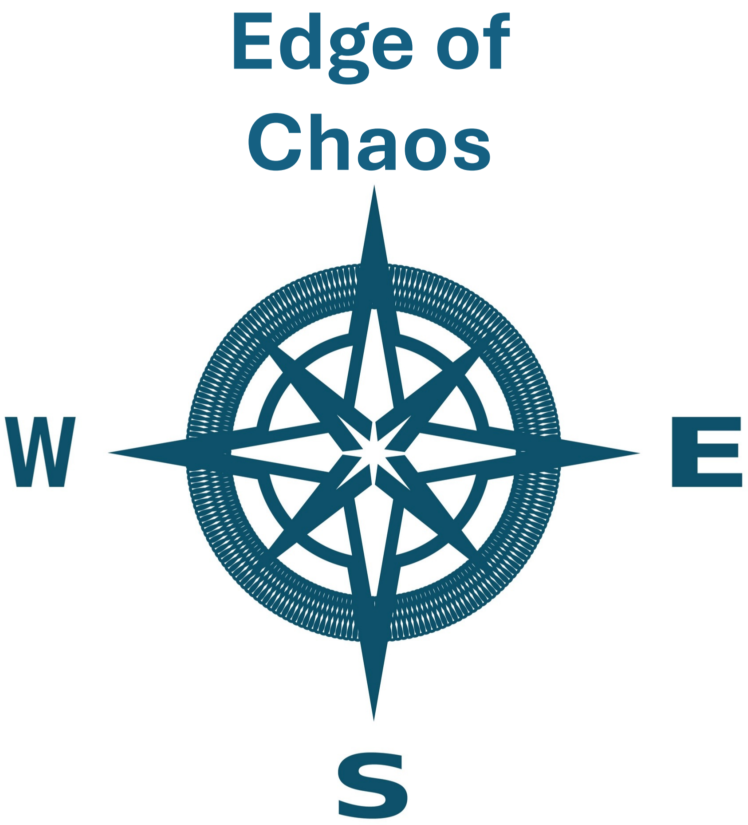
Data Augmentation

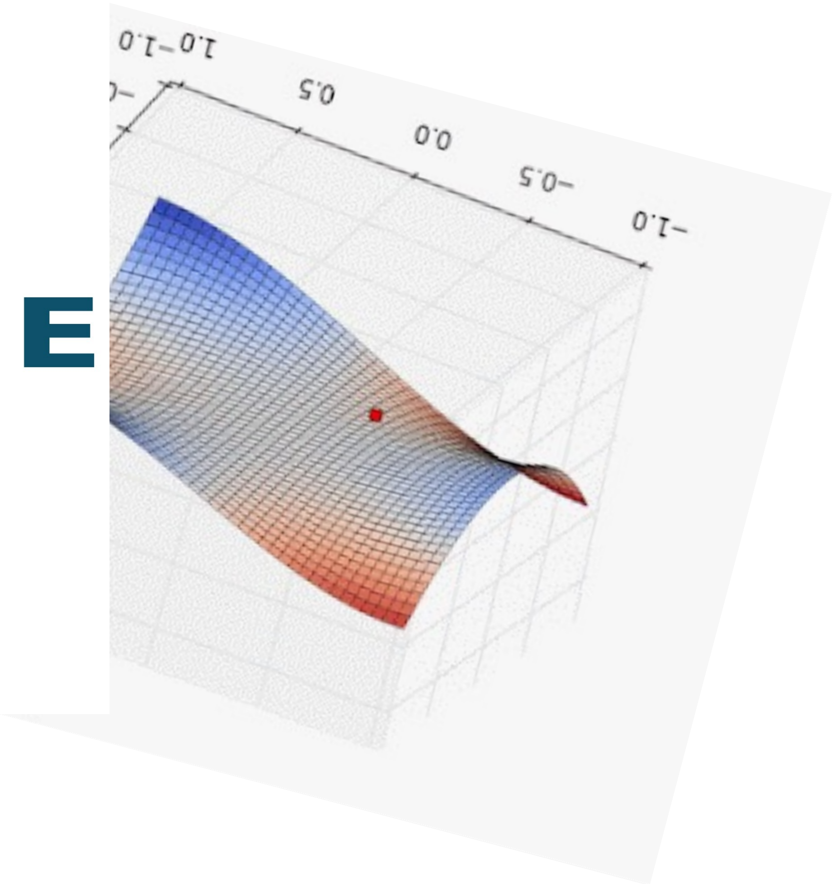
……



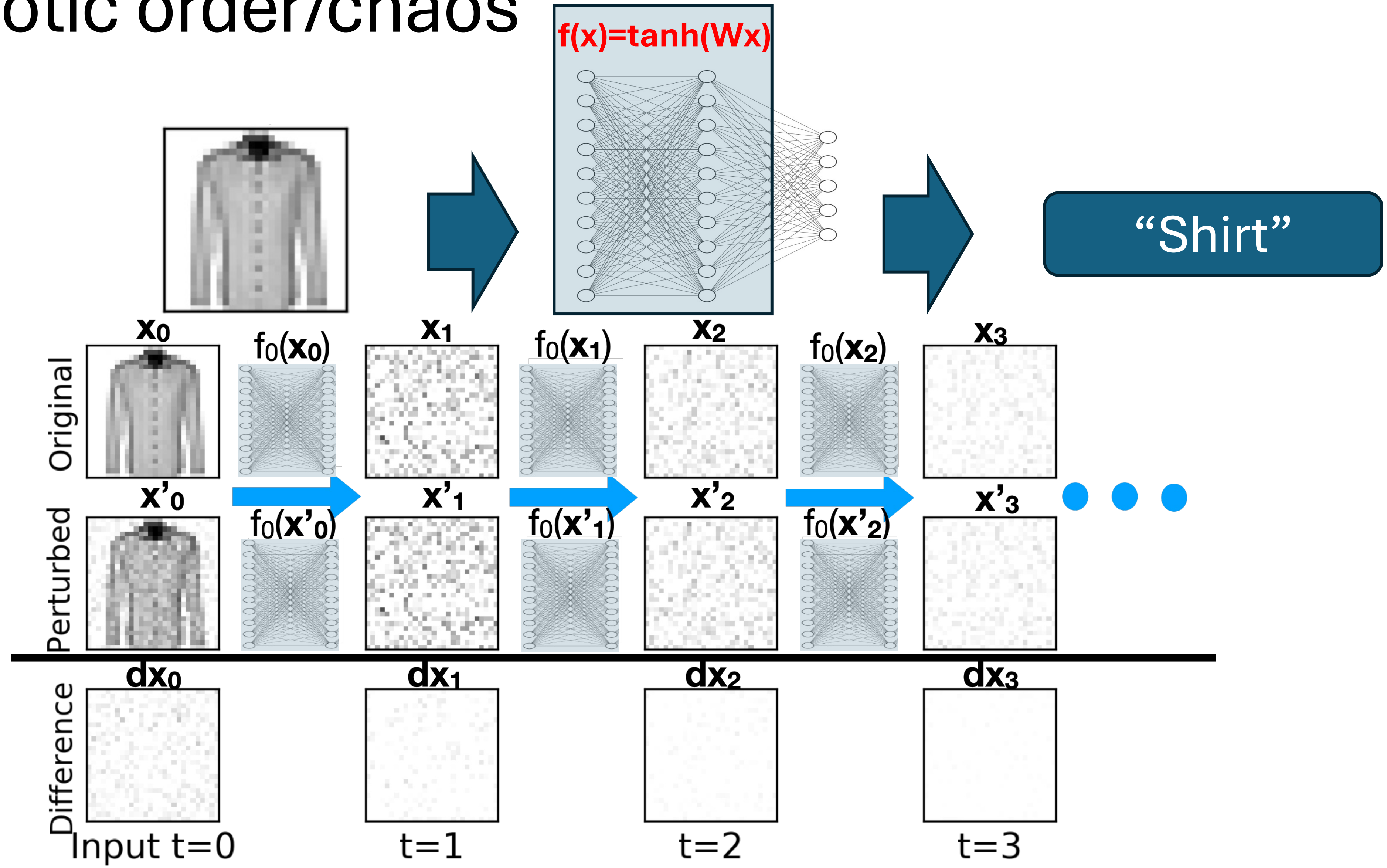
Feng Ling et. al., Optimal Machine Intelligence at the edge of Chaos, arXiv preprint arXiv:1909.05176, 2020

Goal: Train the NN to "***stay***" at edge of chaos



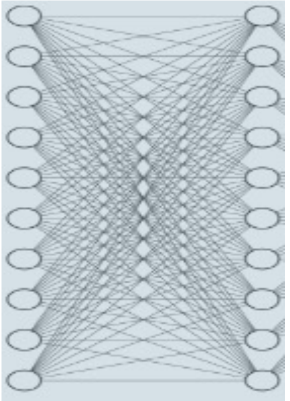


Subject: Single hidden layer feedforward NN
- Asymptotic order/chaos



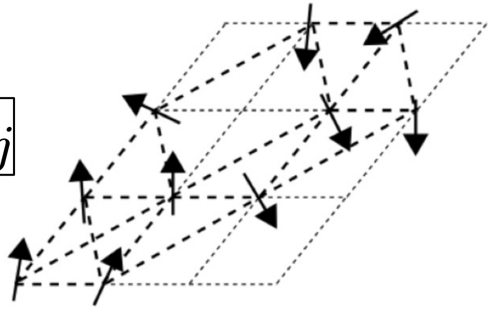
Chaotic phase = SK Spin glass phase

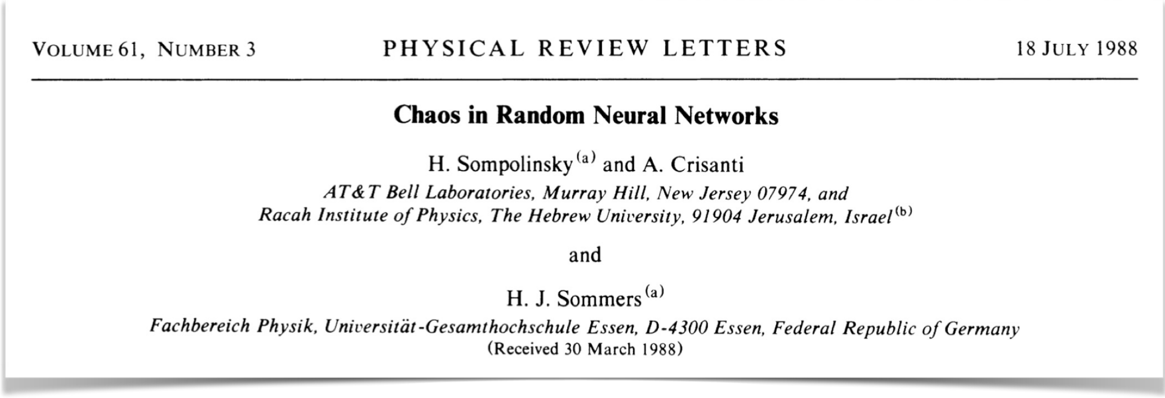
FC with tanh activation



$W_{ij} \to J_{ij}$

SK spin glass model

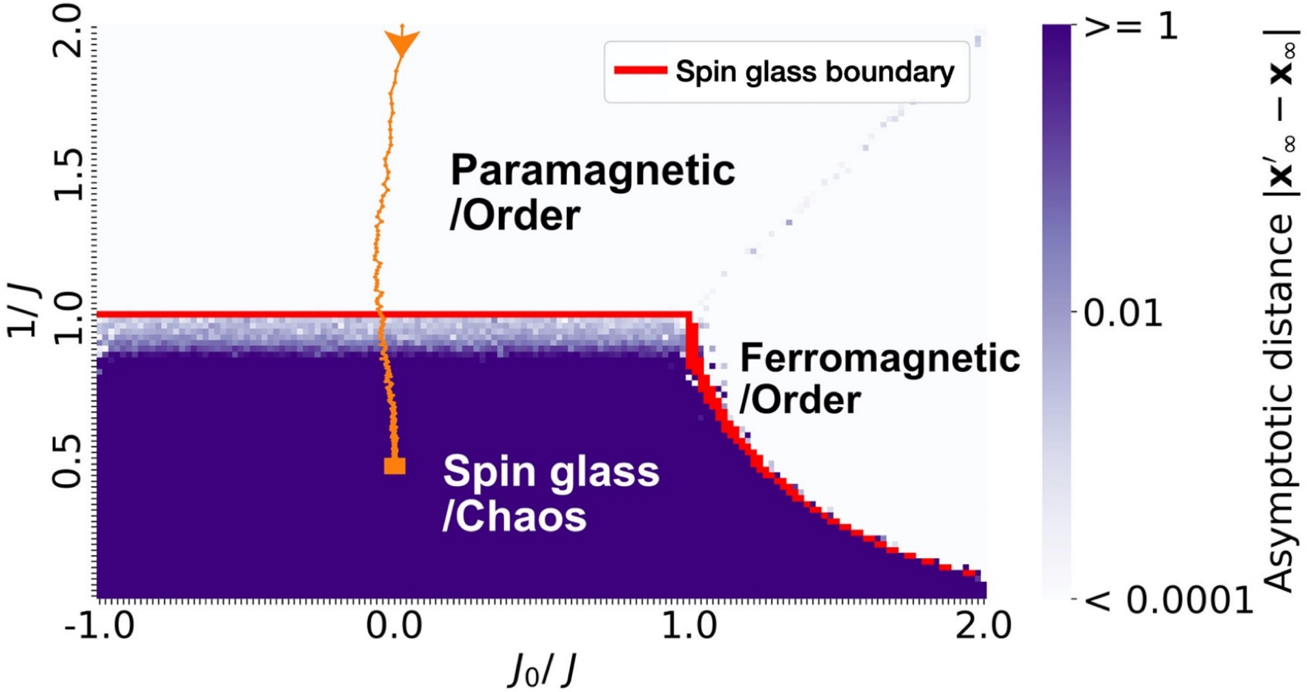




VOLUME 61, NUMBER 3 PHYSICAL REVIEW LETTERS 18 JULY 1988

Chaos in Random Neural Networks

H. Sompolinsky[a] and A. Crisanti
AT&T Bell Laboratories, Murray Hill, New Jersey 07974, and
Racah Institute of Physics, The Hebrew University, 91904 Jerusalem, Israel[b]
and
H. J. Sommers[a]
Fachbereich Physik, Universität-Gesamthochschule Essen, D-4300 Essen, Federal Republic of Germany
(Received 30 March 1988)



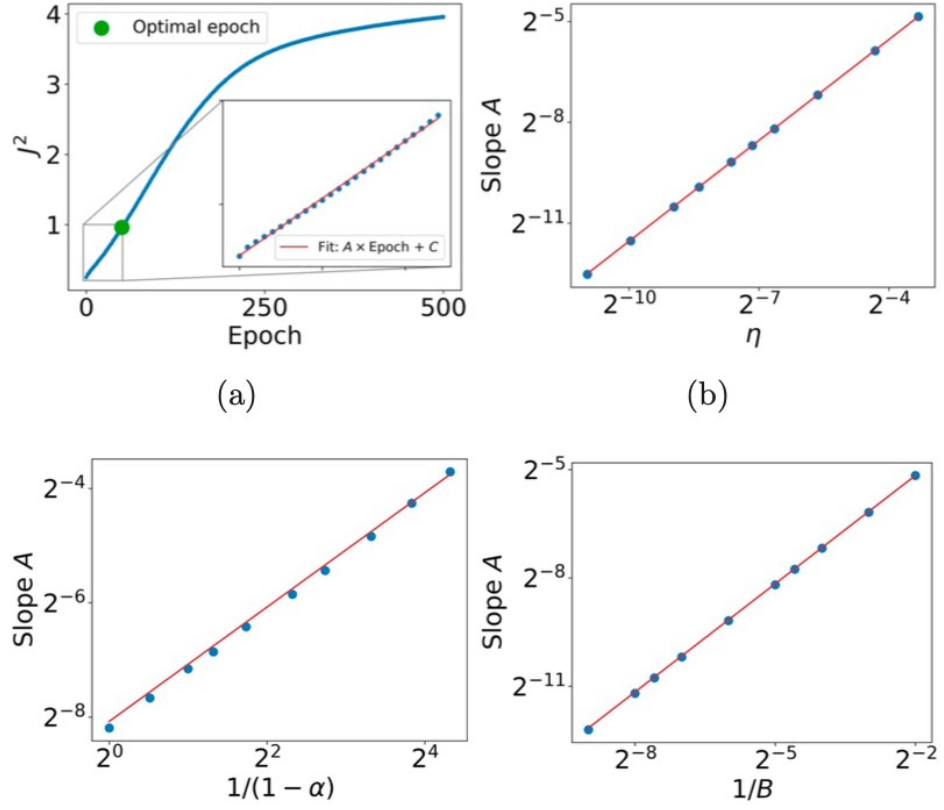
Scaling of hyperparameters – in order phase

η : Learning rate (Step size in gradient)

B : Batch size (# of samples in each gradient computation)

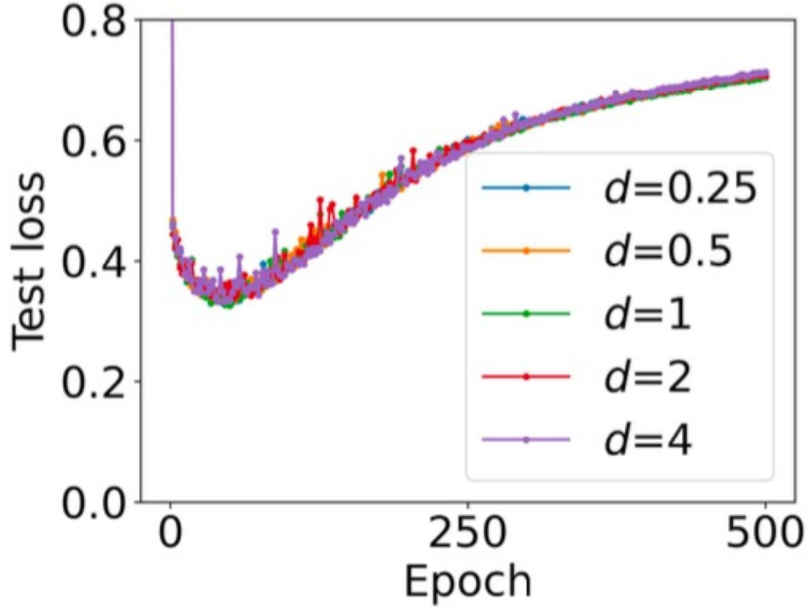
1-α : Momentum (memory of previous gradients)

J^2: normalized variance of weight matrix



(a) (b) (c) (d)

$$d = \frac{\eta}{(1-\alpha)B}$$



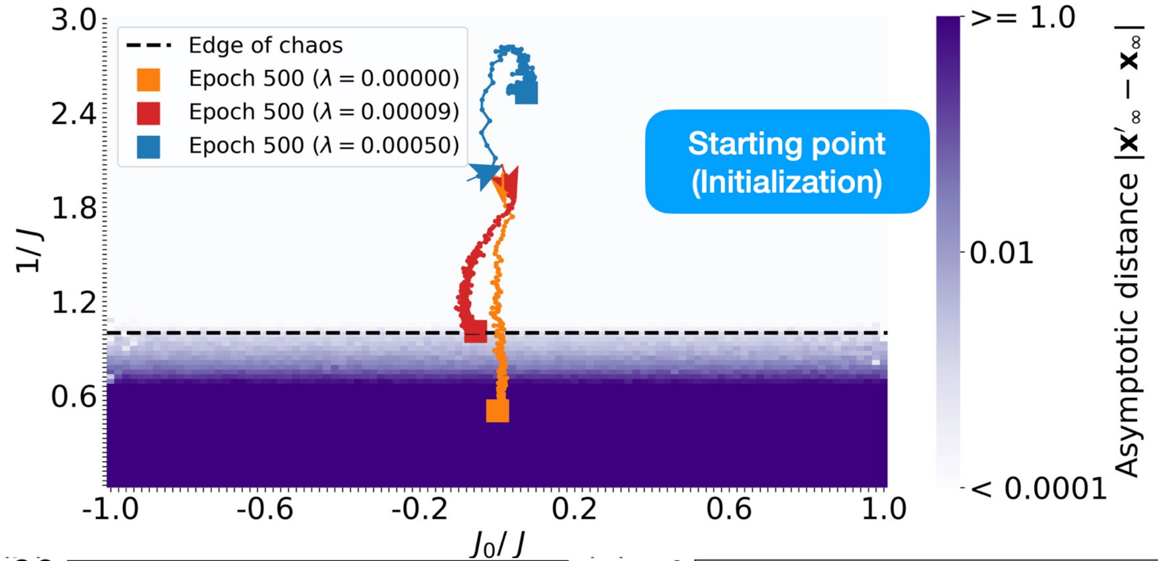
Use linear scaling to control edge of chaos - Optimal L2 regularization (weight decay)

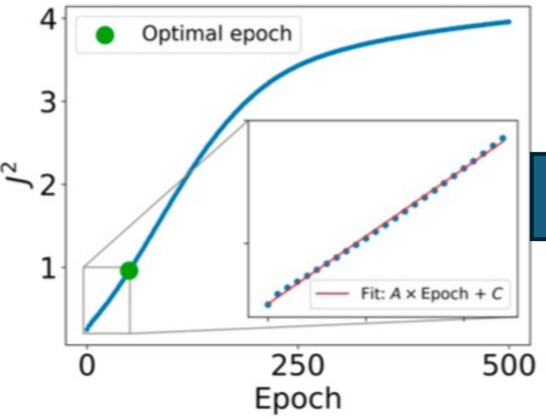
$$\Delta w_t = v_t = \alpha v_{t-1} - \eta g_t$$

Weight decay

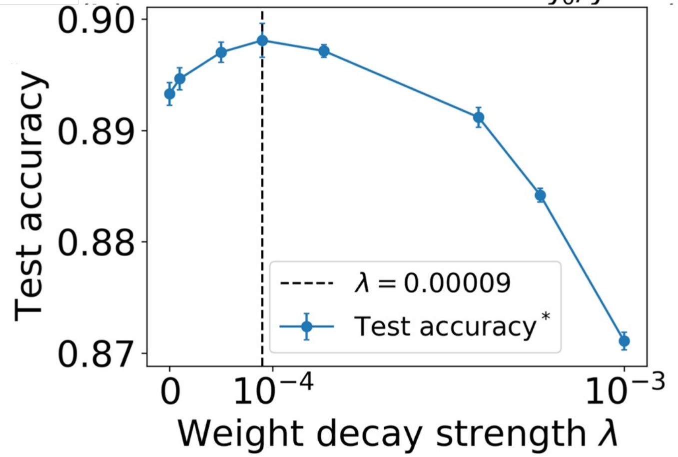
$$\Delta w_t = v_t = \alpha v_{t-1} - \eta(g_t + 2\lambda w_t)$$

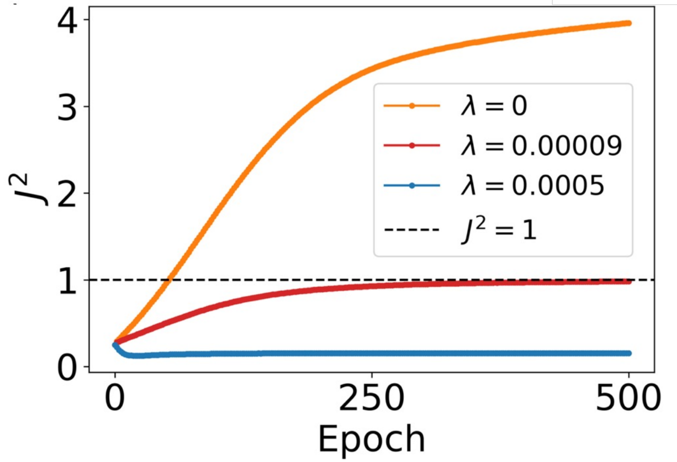
λ: hyperparameter to tune decay strength





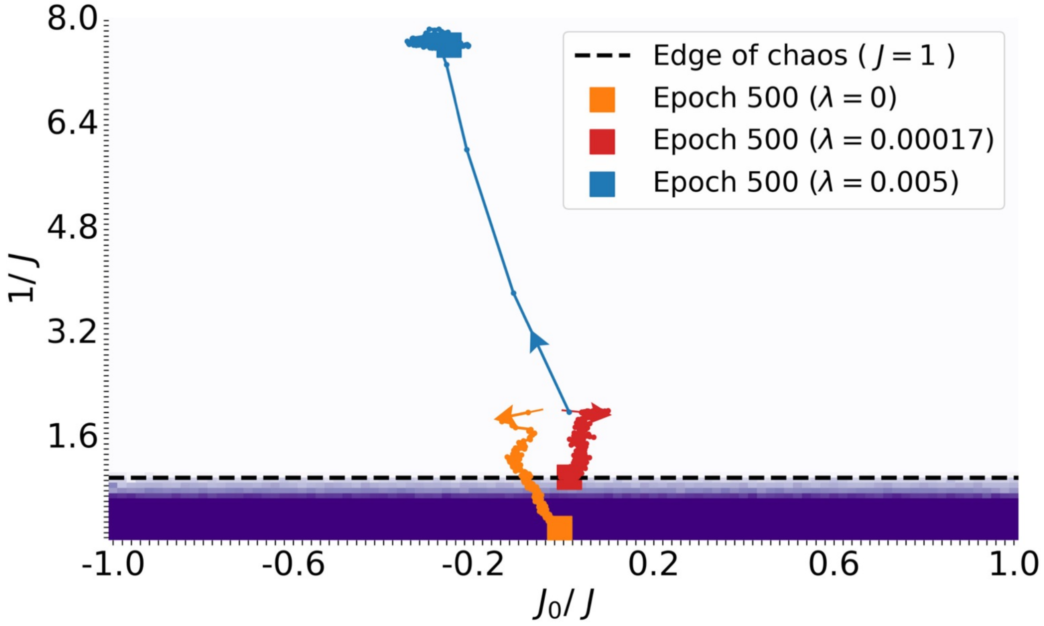
$$\lambda \geq \frac{(1-\alpha)B}{\eta} \cdot \frac{A}{4S} \approx 0.00005$$





Robust training at the edge of chaos - against 20% label noise (shuffled)

The NN automatically avoids fitting wrong labels and gives best test accuracies



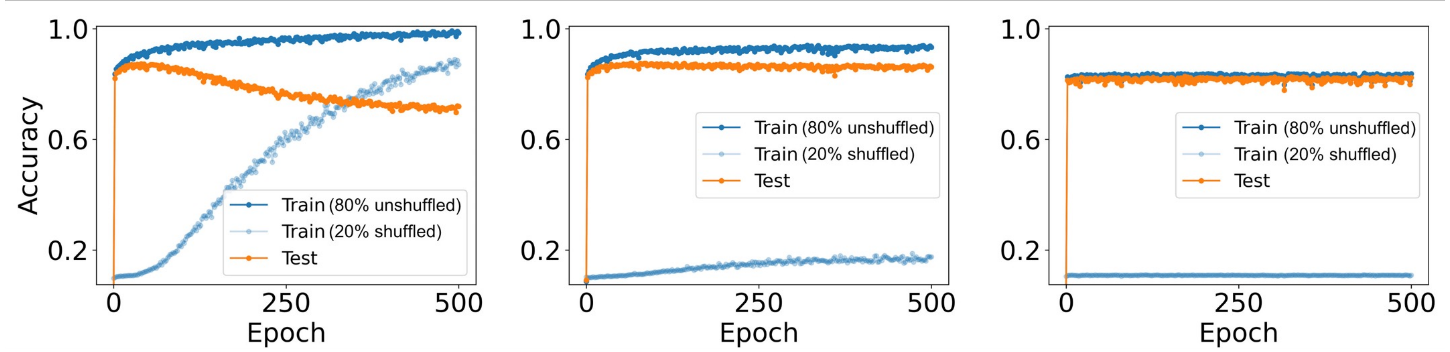
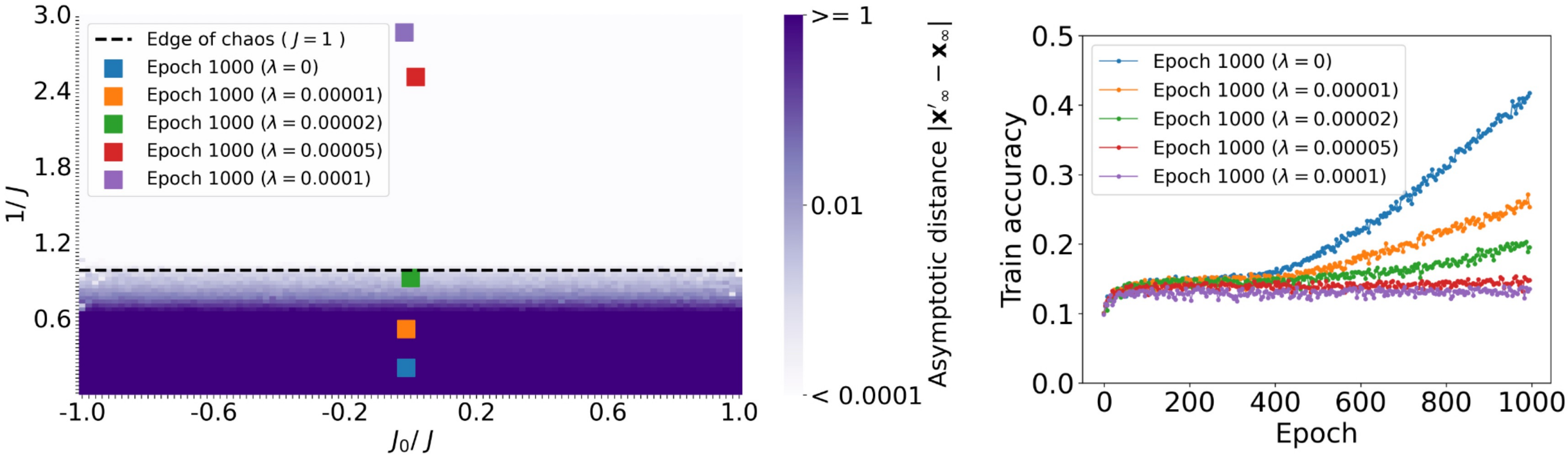


Table 1. Accuracy trained on Fashion-MNIST data with 20% noisy labels after 500 epochs.

	Accuracy		
Model phase	Train (unshuffled 80%)	Train (shuffled 20%)	Test
Order ($\lambda = 0.005$)	82.6%	10.8%	81.4%
Edge of chaos ($\lambda = 0.00017$)	92.7%	16.3%	**86.0%**
Chaos ($\lambda = 0$)	98.4%	87.1%	71.9%

Robust training at the edge of chaos (extreme)
- against 100% label noise (shuffled)

The NN refuses to fit noisy training data in the order phase
It fits more noisy training data as it gets more chaos



Summary

- Even for feedforward (and shallow) networks, asymptotic edge of chaos during training leads to optimal model performances.
- It also leads to robustness against label noise.
- To achieve optimal performance and robustness:
  - Start in the order phase of the model
  - Make the model to stay at edge of chaos while exploring optimal weight configurations – it will not overfit or underfit!
- Challenges:
  - Complex NN architectures (like LLM) likely do not permit analytical solution for edge of chaos
  - 'Engineering' solutions are needed to control the model to stay at the edge of chaos.